Санкт-Петербургский государственный университет

Математическое обеспечение и администрирование информационных систем
Информационно-аналитические системы

Минаев Никита Михайлович

BigData для Умного города. Способы применения

Бакалаврская работа

Научный руководитель:
к. ф.-м. н., доцент Графеева Н. Г.

Рецензент:
Калугин Д. И.

Санкт-Петербург
2017

SAINT-PETERSBURG STATE UNIVERSITY

Software and Administration of Information Systems
Analytical Information Systems

Nikita Minaev

Big Data for Smart city. Method of application

Bachelor's Thesis

Scientific supervisor:
associate professor Natalia Grafeeva

Reviewer:
Dmitry Kalugin

Saint-Petersburg
2017

Оглавление

Введение и постановка задачи

BigData - термин, используемый для обозначения экспоненциально растущего объема доступных данных [8]. Однако, BigData относится не только к самому объему информации, но также к “скорости данных”, что подразумевает потоки данных, которые поступают и обрабатываются в реальном времени, и к различным источникам этих данных.

В современном мире развитие технологий, связанных с обработкой больших данных, позволяет делать определенные предсказания и автоматически решать задачи планирования в самых разных областях. Одним из известных примеров таких областей является ритейл [9], но аналогичные подходы можно использовать для решения задач Умного города.

В данной работе рассмотрены способы использования BigData для решения задач Умного города и приведена реализация одного из них - определения и предсказывания событий и числа их участников. Предложено решение данной задачи с использованием большого объема данных, получаемого с помощью социальной сети ”ВКонтакте”. Для этого разработаны инструменты, позволяющие получать наборы таких данных, выбирать из них наиболее актуальные, визуализировать и анализировать их.

1. Обзор способов применения BigData для Умного города

На текущий момент исследователями рассмотрено множество способов применения BigData для решения задач Умного города. Далее будут приведены некоторые сферы жизни города и подходы, которые в них используются для улучшения функционирования городских служб на основе анализа данных.

1.1. Применения BigData для обеспечения безопасности

В связи с широким распространением камер видеонаблюдения в городах, поток данных с них предоставляет большие возможности для анализа с целью обнаружения нетипичной активности. Такой анализ успешно используется для борьбы с преступностью [22].

Помимо камер видеонаблюдения, существуют способы получения данных для обеспечения безопасности из социальных сетей. Примером является исследование употребления алкоголя городскими жителями и анализ их настроения для обеспечения общественной безопасности [13].

1.2. Применения BigData в здравоохранении

Еще одним применением BigData, связанным с концепцией Умного города является сфера здравоохранения. Источиками данных в этой сфере служат как датчики, оценивающие состояние пациентов медицинских учреждений, так и датчики, встроенные в смартфоны, умные часы и прочие составляющие умной одежды (wearable technology [16]). Анализ таких данных в реальном времени позволяет уведомлять пользователей (пациентов и врачей) об изменениях состояия человека, на основании которых можно предсказывать необходимость скорой госпитализации, а также избегать излишних вызовов служб скорой помощи, совершая необходимые медицинские манипуляции заранее. Также

результаты подобного анализа делают возможным распознавание заболевания на ранних стадиях, а в процессе лечения становятся одним из факторов, влияющих на выбор способов медицинского вмешательства [10].

Другим применением анализа больших данных в этой сфере является распознование мошенничества. Применение таких методов анализа к записям о пациентах и об оплате ими услуг позволяет обнаруживать различные виды мошенничества, например ситуации, в которых обнаруживается, что в соответствии с документами, один пациент в одно и то же время находится в различных клиниках [10].

1.3. Применения BigData в сфере электроснабжения

Методы анализа больших данных также имеют применение в сфере электроснабжения и ключевым понятием при этом являются умные сети электроснабжения (smart grid), которые используют данные об энергопотреблении для оптимизации использования электричества. С ростом доли энергии, получаемой от солнечных батарей и ветряных электростанций, растет также необходимость в предсказании их продуктивности, в силу ее непостоянности. Для этого используется анализ погодных данных [4].

1.4. Применения BigData для организации дорожного движения

Большое число способов применения BigData связано с анализом трафика, прогнозированием дорожных пробок. В частности, система московского Центра организации дорожного движения использует такие подходы для оптимизации транспортных потоков. На основании данных о движении транспорта эта система информирует население о пробках и управляет подключенными к ней светофорами [22], [1].

Существуют и другие применения BigData, связанные с управлени-

ем дорожным движением. Примерами являются: устранение неровностей дорожного полотна на основе данных акселерометров телефонов жителей, помощь водителям с нахождением места для парковки исходя из загруженности места назначения на основе датчиков [18].

Для анализа трафика важной задачей является обнаружение событий (event detection), которые могут сильно повлиять на дорожную обстановку. Эта задача заключается в поиске городских событий, таких как фестивали, концерты, спортивные мероприятия и т.д., которые привлекают большое число людей [7]. Ее решение может помочь планированию городской инфраструктуры и ее подготовки к этим событиям. Далее в данной работе будет более подробно рассмотрена эта задача, приведен обзор существущих решений с использованием данных, полученных из социальных сетей и предложена реализация системы для обнаружения событий на основе данных из социальной сети "ВКонтакте". Также будут описаны результаты применения алгоритмов анализа временных рядов для определения ожидаемого числа участников.

2. Обзор существующих решений

Примеры использования BigData для определения мест скопления людей на основе данных социальных сетей и сопоставления полученных локаций с данными о городском транспорте описаны в [12], а также в [14], где аналогичные подходы используются для определения происходящих в городе событий. Практическая польза от использования социальных сетей для предсказания скоплений людей также упоминается в статье об инфраструктуре Fujitsu SPATIOWL [17].

Авторы исследования [14] рассматривают город Барселону и определяют 20 событий, оказавших самое сильное влияние на социальные сети за рассматриваемый период. 12 из них оказались матчами ФК "Барселона", которые посетило большое число людей (до 98000). На основании исследования 100 самых популярных событий, авторы указывают на корреляцию между данными полученными ими из социальных сетей и реальным числом участников мероприятия.

1	FCB vs Madrid	11	3 nearby concerts
2	FCB vs Elche	12	Airport
3	FCB vs Malaga	13	Michael Buble concert
4	FCB vs RCDE	14	Arctic Monkeys concert
5	FCB vs Milan	15	New Year @ Park Guell
6	FCB vs Granada	16	Airport
7	FCB vs Valencia	17	Bruno Mars concert
8	FCB vs Villareal	18	FCB vs Efes (Basketball)
9	FCB vs Celtic	19	FCB vs Real Sociedad
10	FCB vs Cartagena	20	Airport

Рис. 1: 20 событий, сильнее всего повлиявших на социальные сети в период, в который проводились измерения

В этих статьях, как и в других, связанных с определением событий на основе данных социальных сетей (например, [11]), основным источником данных является Twitter. К нему также иногда добавляют Foursquare и Instagram с целью увеличения размеров выборки и борьбы с выбросами. Социальная сеть "ВКонтакте" собирает гораздо больше информации о своих пользователях, чем упомянутые выше ресурсы и предоставляет к ней доступ посредством VK API. Таким образом, использование этой сети для определения событий, позволило бы полу-

чить множество различных данных об участниках каждого из них. На основании этих данных можно в дальнейшем получить больше информации о самом событии.

Второй причиной для выбора "ВКонтакте" является его популярность в России, поэтому данные полученные с помощью VK API лучше подходят для предсказания событий в российских городах.

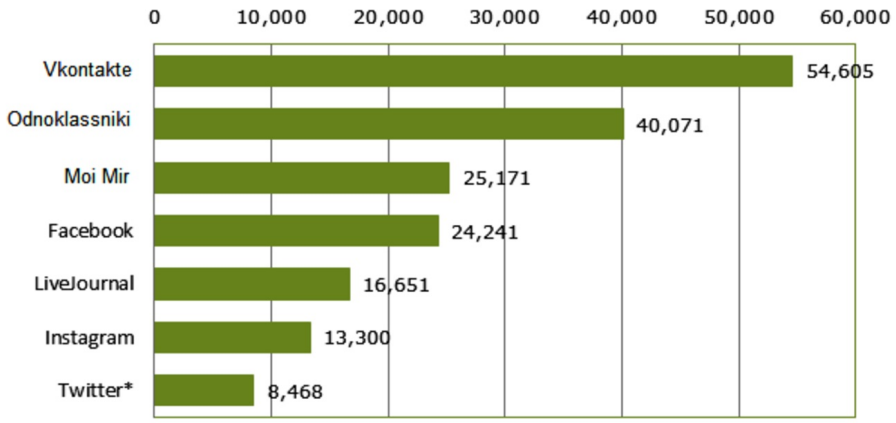

Рис. 2: Сравнение аудитории социальных сетей в России (тыс. человек в месяц) [15]

3. Описание получаемых данных

Все данные в этой работе получаются с помощью VK API [21]. Целью обращения к VK API является получение записей о событиях. В сети "ВКонтакте" события являются частными случаями групп и нет способа заранее узнать, представляет ли данная группа событие, кроме как получив данные по идентификатору группы. Поэтому, первой полученной выборкой был набор групп, полученных перебором идентификаторов до некоторого значения. При этом, в выборку добавлялись только те группы, для которых были указаны географические координаты.

Эта выборка сама по себе не является информативной в рамках данной работы, но при ее рассмотрении стало понятно, что среди групп регулярно встречаются такие, которые представляют собой события. Поэтому алгоритм был модифицирован, чтобы сохранять только группы, тип которых указан как *event*. При этом сохраненные события не всегда идут в хронологическом порядке, и актуальные события могут соседствовать с давно завершившимися.

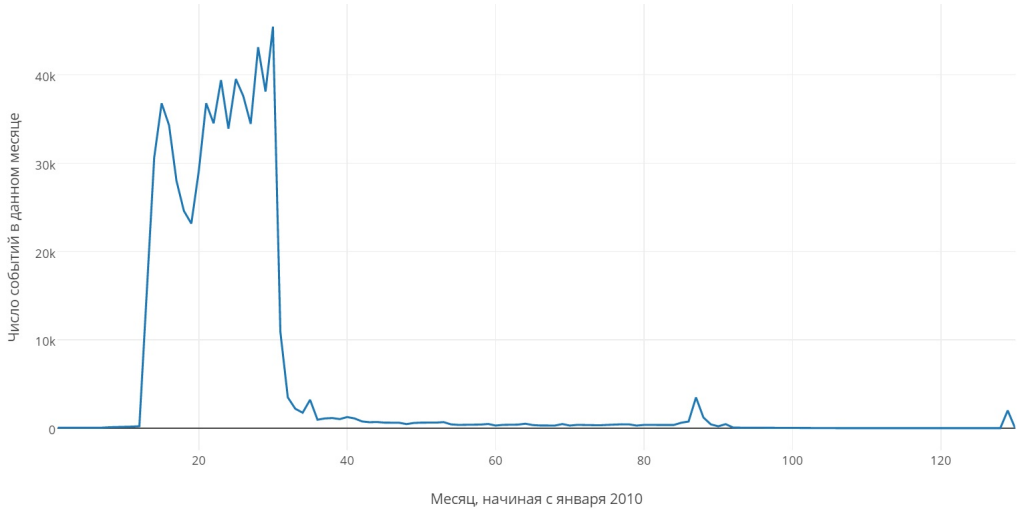

Рис. 3: График, показывающий число сохраненных событий из первой выборки с событиями

Помимо этих данных о событиях, были также получены данные о количестве пользователей, решивших их посетить, полученные в разные моменты времени. На их основе можно делать предположения об

ожидаемом количестве участников. Эти данные были получены также с помощью регулярных запросов к VK API с целью узнать количество участников, отметившихся в группе того или иного события.

Когда во "ВКонтакте" создается событие, люди планирующие его посетить отмечают, что придут туда. Таким образом, как правило, число участников события растет с нуля, в момент его создания, до некоторого числа, близкого к числу посетителей, в момент когда событие наступило. Поэтому можно говорить о том, что в этих данных присутствует тренд.

4. Реализация алгоритмов работы с данными

Для получения и анализа данных в рамках данной работы были реализованы алгоритмы на языке PHP с использованием баз данных под управлением MySQL. Далее они будут рассмотрены подробнее.

4.1. Реализация алгоритма получения данных

Чтобы получать описанные выше данные, на хостинге запущен выполняющийся по таймеру скрипт. Этот скрипт отправляет запросы к "ВКонтакте" посредством VK API и получает ответ в формате JSON. Затем представляющие интерес данные из полученного ответа (а именно, страна, город, широта и долгота, название группы, количество участников для групп и, помимо этого, время начала и окончания для событий) сохраняются в базу данных, расположенную на хостинге.

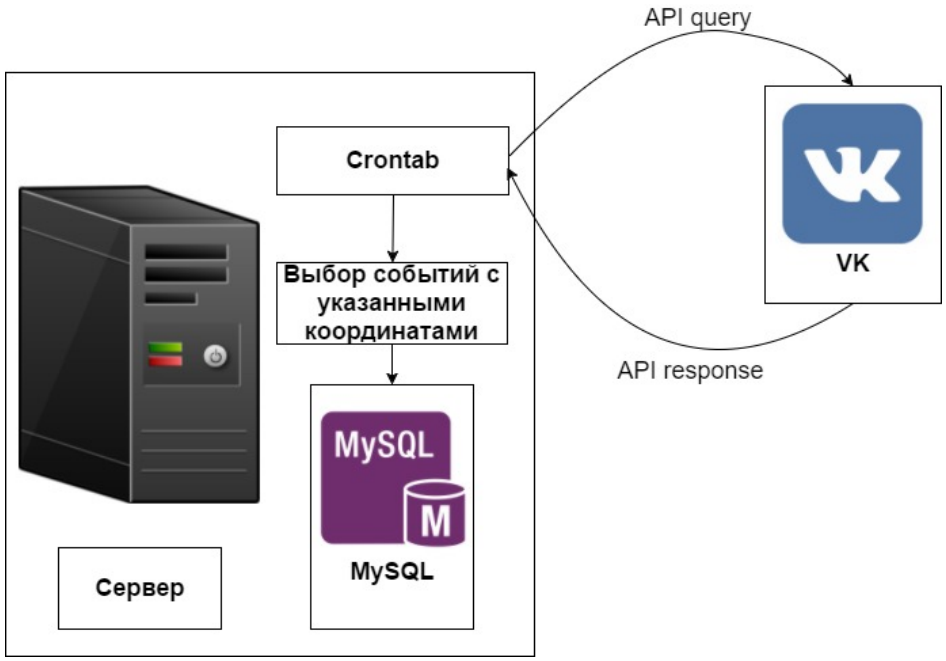

Рис. 4: Схема получения выборки событий

При получении первой выборки было просмотрено 10019999 групп и из них сохранены 65914 - те, для которых были указаны географические координаты. Затем, для получения выборки с событиями было

просмотрено 42762249 групп и из них сохранено 688975 событий. Основным ограничением этого процесса сбора информации является скорость обработки запросов к VK API - она ограничена 5 запросами в секунду.

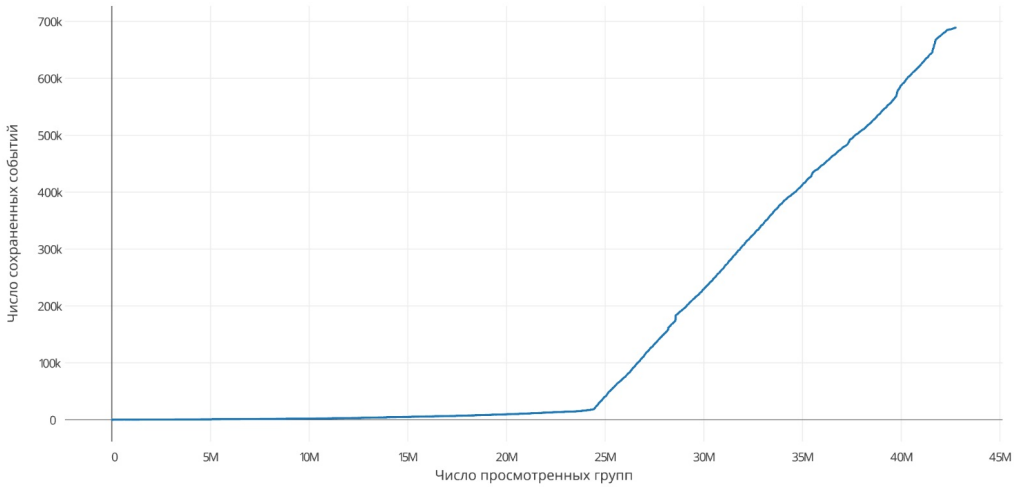

Рис. 5: График зависимости числа сохраненных событий от числа просмотренных групп

В ходе получения данных о динамике количества пользователей, отметившихся в группе события, первым этапом были выбраны исследуемые события: которые должны были произойти в мае 2017 года и в группах которым отметилось не менее 1000 участников. Для этих событий производились регулярные измерения числа отметившихся участников. Это также было реализовано с помощью скрипта, исполняющегося по таймеру и записывающего результаты своей работы в базу данных. После этого был рассмотрен прирост количества участников с 02.05.2017 по 19.05.2017.

Событие	Кол-во участников 02.05.2017 19:40	Кол-во участников 19.05.2017 05:40	Изменение
“Больше красок! Фестиваль холи Пермь 27-28 мая”	2171	2527	16.4%
“Фестиваль Водных фонариков – Санкт-Петербург“	9959	10549	5.9%
“Фестиваль "Красота на Волге"”	1381	1458	5.6%
“МОТО Весна VII с 19 по 21 мая 2017 года“	2258	2361	4.6%
“Московский Велопарад“	16526	17270	4.5%
“20-21 мая: Интеллигентная барахолка”	14188	14563	2.6%

Рис. 6: Таблица событий с наибольшим приростом числа участников с 02.05.2017 по 19.05.2017

4.2. Использованные алгоритмы анализа данных

Для планирования городской инфраструктуры может быть полезно не только определять ожидаемые события, но и выделять группы событий, которые произойдут примерно в одно и то же время, недалеко друг от друга. Для нахождения таких групп, к полученным данных был применен алгоритм кластеризации пространственных данных с присутствием шума DBSCAN [2]. Данный алгоритм был выбран из-за его способности работать с данными с шумом, а также в связи с тем, что ему не нужно подавать на вход количество кластеров.

Для прогнозирования числа участников того или иного мероприятия были использованы методы анализа временных рядов. Исходными данными для них являлись замеры количества пользователей, отметившихся в группах событий, полученные как указано выше. В первую очередь был рассмотрен метод *Simple exponential smoothing (SES)* [5]. Результаты, полученные с помощью SES оказались далеки от реальных данных. Это объясняется тем, что SES плохо обрабатывает данные, в которых присутствует тренд [6]. Поэтому был использована модель *Double exponential smoothing (DES)*, подходящая для работы с такаими данными [3].

5. Полученные результаты

Полученные данные были визуализированы с помощью Google Maps API [20]. Каждый маркер соответствует группе из ”ВКонтакте” и указывает ее название и количество участников.

Первой была визуализирована выборка, состоящая из групп. При этом подтвердилось предположение о том, что она не имеет практического применения в рамках данной задачи, т.к. большинство групп не представляют какого-либо события. Тем не менее, визуализация этой выборки наглядно продемонстрировала большой географический разброс полученных данных.

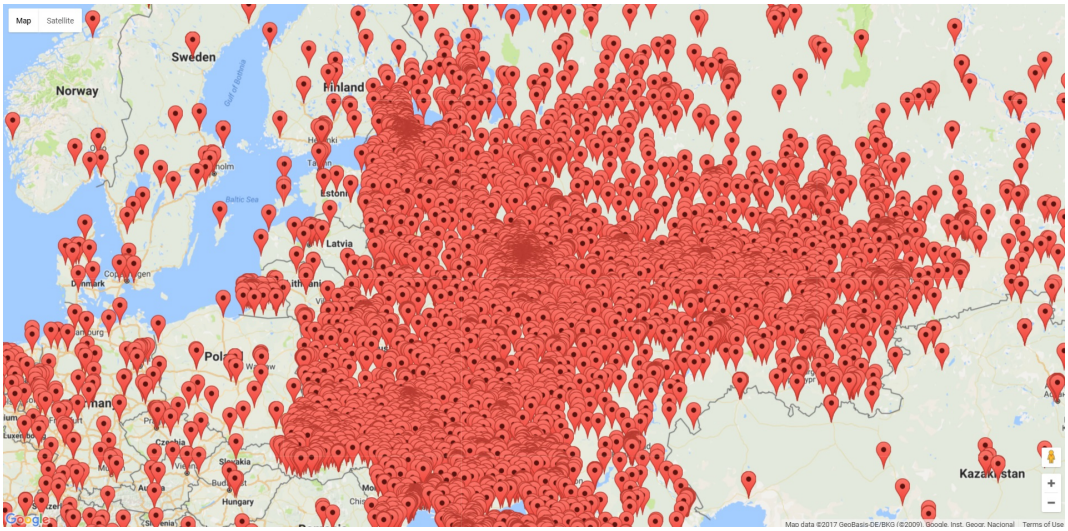

Рис. 7: Выборка состоящая из различных групп ”ВКонтакте”

Затем были визуализированы различные подвыборки данных о событиях. Был реализован интерфейс, позволяющий быстро установить период, в который происходили или будут происходить рассматриваемые события, а также минимальное число их участников.

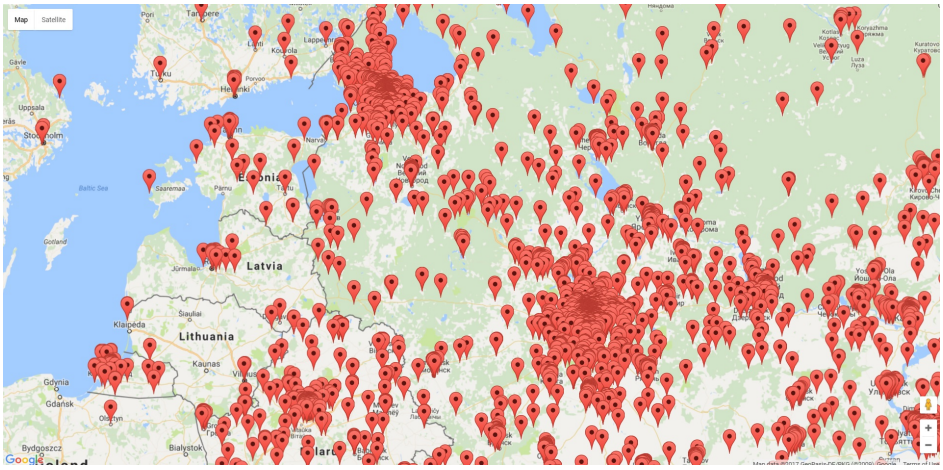

Рис. 8: Все события, произошедшие с 2013 года

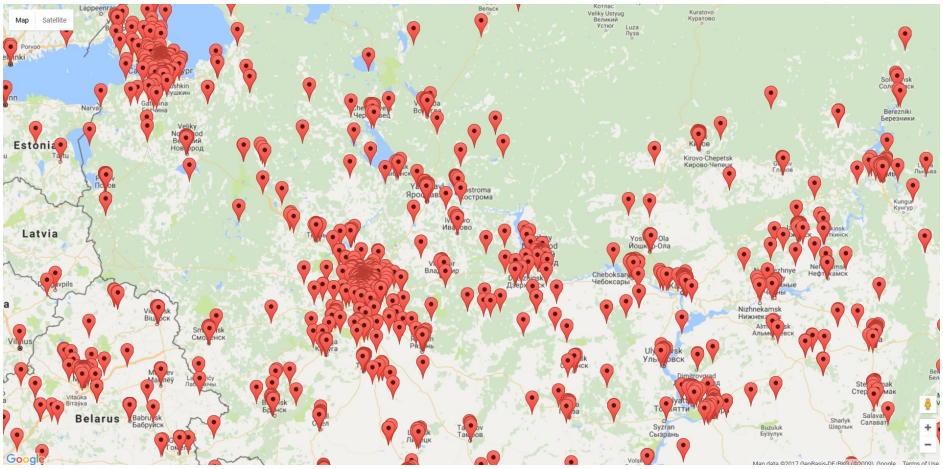

Рис. 9: Все события, которые произошли или произойдут в 2017 году

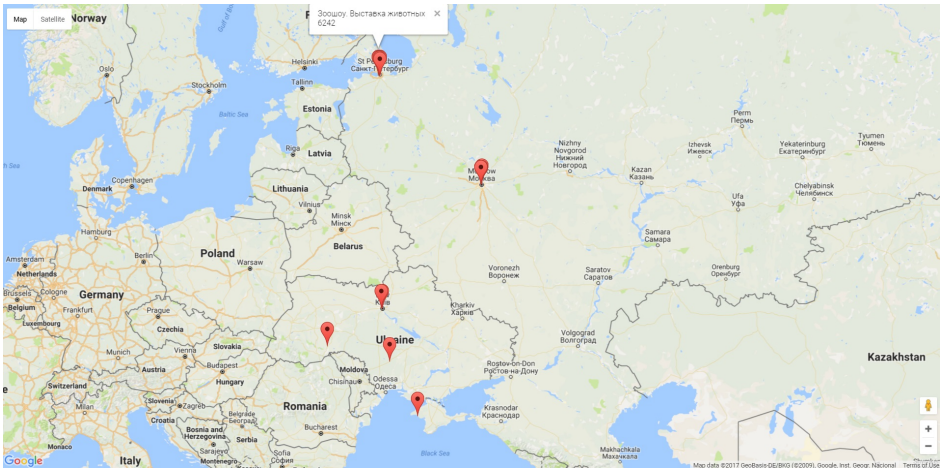

Рис. 10: Все события, которые произошли в марте 2017 года с не менее чем 5000 участниками

Из полученных изображений видно, что как и ожидалось, больше всего массовых мероприятий проходят в крупных городах. Это подтверждается результатами проведенной кластеризации событий.

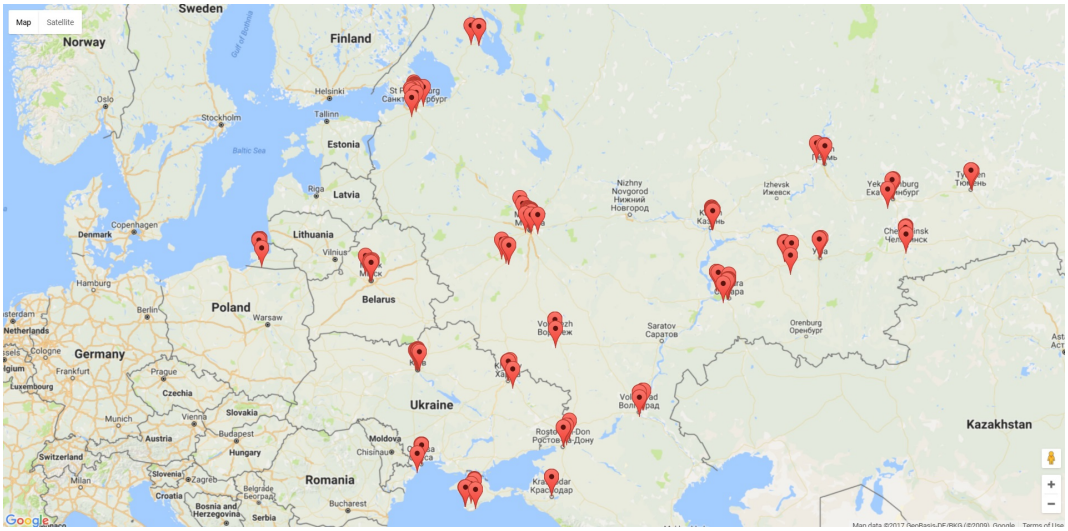

Рис. 11: Кластеризация событий, которые произойдут с 1 по 10 мая 2017 года

С помощью метода Double exponential smoothing, были получены прогнозы для числа участников некоторых событий. Например, в 00:00 19.05.2017 (т.е. перед началом мероприятия) реальное число отметившихся участников события “MOTO Весна VII с 19 по 21 мая 2017 года“ [19] составило 2361 человек, а спрогнозированное - 2358.

Также этот метод был применен к событию с самым большим приростом пользователей за период с 01.05.2017 по 06.05.2017 - ”Фестиваль Водных фонариков – Санкт-Петербург” [23]. Для этого события модель DES также показала правдоподобные результаты, а модель SES - нет.

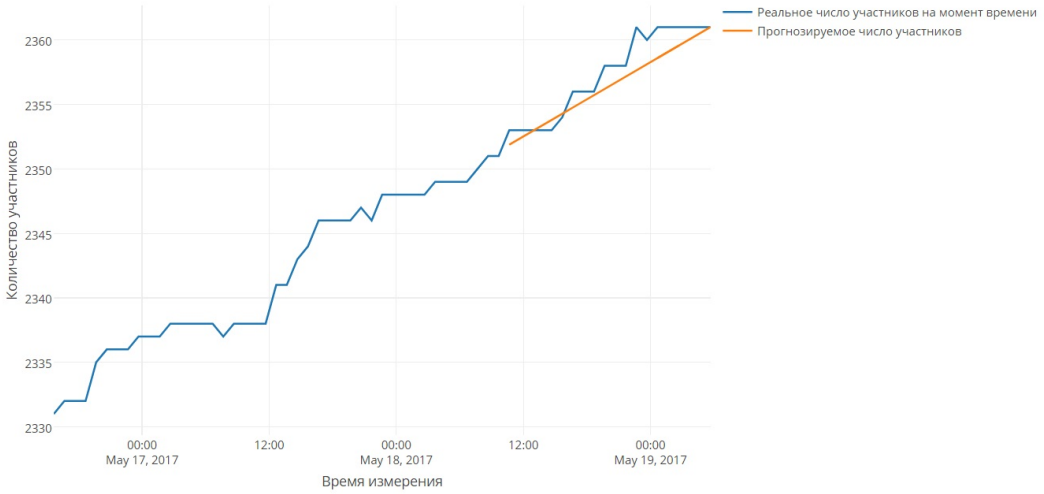

Рис. 12: Реальное и спрогнозированное число участников для события “МОТО Весна VII с 19 по 21 мая 2017 года“

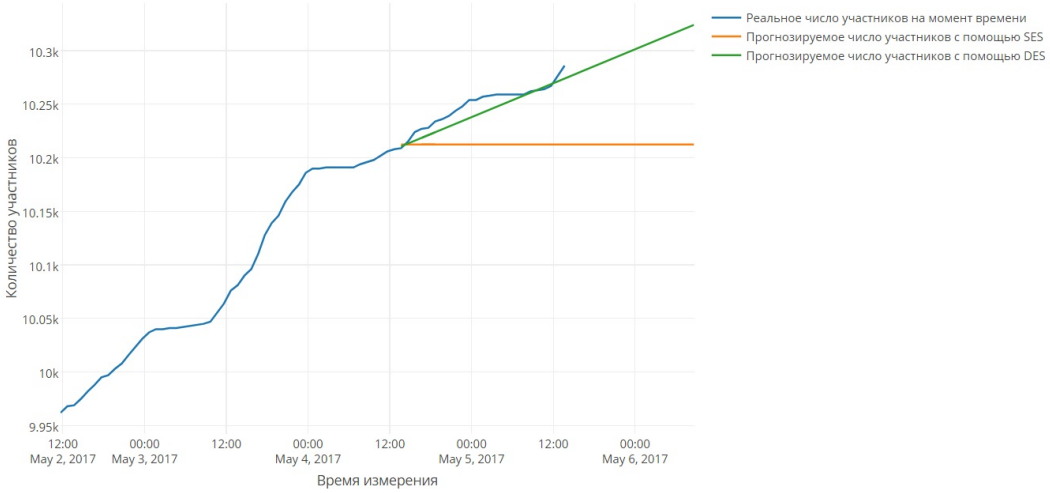

Рис. 13: Реальное и спрогнозированное число участников для события “Фестиваль Водных фонариков – Санкт-Петербург“

Заключение

В рамках данной работы рассмотрены различные применения BigData для решения задач Умного города. Подробно рассмотрены задача определения и предсказания происходящих в городе событий. Получены и визуализированы наборы данных, представляющих события в социальной сети "ВКонтакте" за различные периоды времени. Созданы и многократно использованы инструменты для дальнейшего расширения этих наборов данных. Также использованы методы анализа данных для кластеризации событий и прогнозирования числа их участников.

Список литературы

[1] Applications of big data to smart cities / Eiman Al Nuaimi, Hind Al Neyadi, Nader Mohamed, Jameela Al-Jaroodi // Applications of big data to smart cities. — 2015. — 15 p. — URL: https://jisajournal.springeropen.com/articles/10.1186/s13174-015-0041-5.

[2] A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise / Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu // Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96). — 1996.

[3] Double Exponential Smoothing // NIST/SEMATECH e-Handbook of Statistical Methods. — 2013. — URL: http://www.itl.nist.gov/div898/handbook/pmc/section4/pmc433.htm (дата обращения: 21.05.2017).

[4] Energy Big Data: A Survey / Hui Jiang, Kun Wang, Yihui Wang et al. // Applications of big data to smart cities. — 2016. — P. 3844–3861. — URL: http://ieeexplore.ieee.org/document/7548112/.

[5] Eva Ostertagova, Oskar Ostertag. The Simple Exponential Smoothing Model // MODELLING OF MECHANICAL AND MECHATRONIC SYSTEMS 2011. The 4th International conference. — 2011. — P. 380–384. — URL: https://www.researchgate.net/publication/256088917_The_Simple_Exponential_Smoothing_Model.

[6] Exponential smoothing // Wikipedia. — 2017. — URL: https://en.wikipedia.org/wiki/Exponential_smoothing (дата обращения: 21.05.2017).

[7] Fine-Grained Urban Event Detection and Characterization Based on Tensor Cofactorization / Longbiao Chen, Jeremie Jakubowicz, Dingqi Yang et al. // IEEE Transactions on Human-Machine

Systems. — 2015. — 12 p. — URL: http://ieeexplore.ieee.org/document/7547355/.

[8] Glossary Gartner IT. What Is Big Data? // Technology Research | Gartner Inc. — 2016. — URL: http://gartner.com/it-glossary/big-data/ (дата обращения: 21.02.2017).

[9] How Big Data Analysis helped increase Walmarts Sales turnover? // DeZyre. — 2017. — URL: https://www.dezyre.com/article/how-big-data-analysis-helped-increase-walmarts-sales-turnover/109 (дата обращения: 22.04.2017).

[10] McDonald Carol. 5 Big Data Trends in Healthcare for 2017 // MapR. — 2017. — URL: https://mapr.com/blog/5-big-data-trends-healthcare-2017/ (дата обращения: 23.04.2017).

[11] Multiscale event detection in social media / Xiaowen Dong, Dimitrios Mavroeidis, Francesco Calabrese, Pascal Frossard // Data Min Knowl Disc. — 2014. — P. 1374–1405.

[12] Smart Cities, Urban Sensing and Big Data: Mining Geo-location in Social Networks / Daniele Sacco, Gianmario Motta, Linlin You et al. // Congresso Nazionale AICA. — 2013. — 10 p.

[13] Social-Cultural Monitoring of Smart Cities Using Big Data Methods Alcohol Consumption and Sentiments / Mengdi Li, Eugene Ch'ng, Boying Li, Shunyi Zhai // 3rd International Conference on Smart Sustainable City and Big Data. — 2015. — 8 p. — URL: https://www.academia.edu/17643106/Social-Cultural_Monitoring_Of_Smart_Cities_Using_Big_Data_Methods_Alcohol_Consumption_And_Sentiments.

[14] Social network data analysis for event detection / Dario Garcia-Gasulla, Sergio Alvarez-Napagao, Arturo Tejeda-Gomez et al. // ECAI. — 2014. — P. 1009–1010. — URL: https://upcommons.upc.edu/bitstream/handle/2117/27743/FAIA263-1009.pdf.

[15] Top social networks in Russia: latest numbers and trends // Russian Search Tips. — 2015. — URL: http://www.russiansearchtips.com/2015/01/top-social-networks-russia-latest-numbers-trend/ (дата обращения: 22.03.2017).

[16] Wearable technology // Wikipedia. — 2017. — URL: https://en.wikipedia.org/wiki/Wearable_technology (дата обращения: 24.04.2017).

[17] «Умные города» и Большие Данные // Блог компании Fujitsu / Хабрахабр. — 2015. — URL: https://habrahabr.ru/company/fujitsu/blog/258925/ (дата обращения: 21.02.2017).

[18] Богданов Александр. Как используется Big Data в больших городах // Downtown. — 2017. — URL: http://downtown.ru/voronezh/technology/8847 (дата обращения: 24.04.2017).

[19] MOTO Весна VII с 19 по 21 мая 2017 года // ВКонтакте. — 2017. — URL: https://vk.com/club34810114 (дата обращения: 22.05.2017).

[20] Начало работы | Google Maps Javascript API. — 2016. — URL: https://developers.google.com/maps/documentation/javascript/tutorial?hl=ru (дата обращения: 21.02.2017).

[21] Разработчикам. — 2016. — URL: https://vk.com/dev (дата обращения: 21.02.2017).

[22] Тишина Юлия. Данные ударят по газам // Российская Газета. — 2016. — URL: https://rg.ru/2016/11/23/sistema-umnyj-gorod-smozhet-uluchshit-zhizn-naseleniia.html (дата обращения: 23.04.2017).

[23] Фестиваль Водных фонариков – Санкт-Петербург // ВКонтакте. — 2017. — URL: https://vk.com/club35597787 (дата обращения: 22.05.2017).